Traffic crash severity prediction using synthesized crash description narratives and large language model (LLM)

Mohammadjavad Bazdar, Md Tufajjal Hossain, Joyoung Lee, Branislav Dimitrijevic

John A. Reif, Jr. Department of Civil and Environmental Engineering

NJDOT BRIIT TechTalk!

May 14, 2025

NJIT
New Jersey Institute of Technology

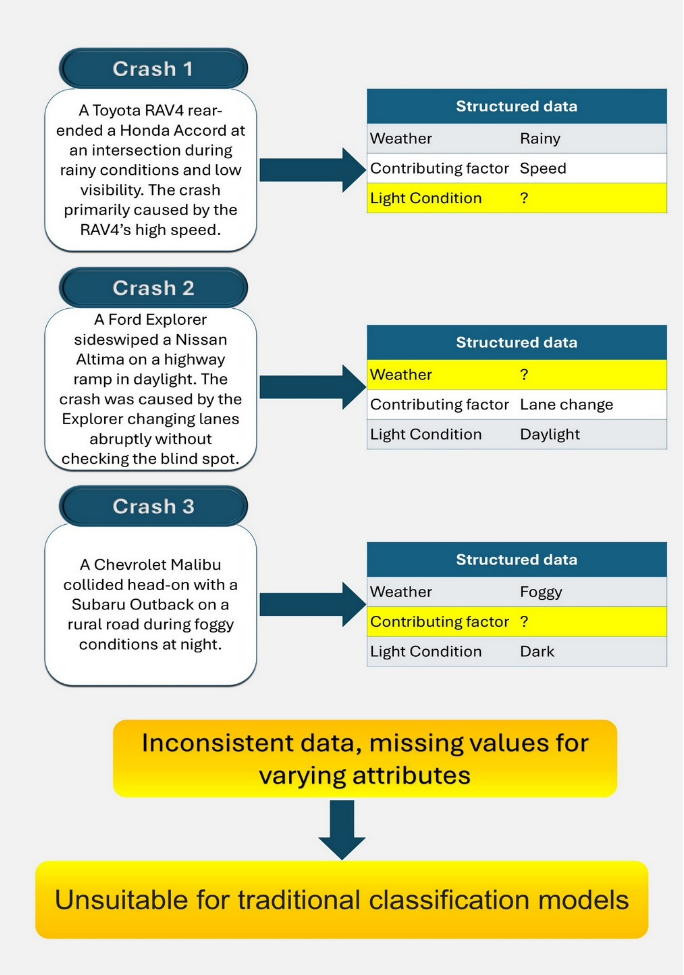
ITSRC
Intelligent Transportation Systems Resource Center

Outline

1. Background
2. Research Objective
3. Methodology & Data
4. Preliminary Results
5. What’s Ahead

Background

- Predicting Crash Severity Matters:
  - **Prevention:** Assess the contributing factors, anticipate when/where the severe crashes might occur.
  - **Faster Emergency Response**: Prioritize high-severity crashes for urgent aid
  - **Efficient Resource Allocation**: Optimize medical and law enforcement dispatch
  - **AV Behavior Modeling:** Train autonomous vehicles to react appropriately
  - **Improved Infrastructure Planning:** Identify hazardous areas for safety upgrades



Crash 1

A Toyota RAV4 rear-ended a Honda Accord at an intersection during rainy conditions and low visibility. The crash primarily caused by the RAV4's high speed.

Structured data	
Weather	Rainy
Contributing factor	Speed
Light Condition	?

Crash 2

A Ford Explorer sideswiped a Nissan Altima on a highway ramp in daylight. The crash was caused by the Explorer changing lanes abruptly without checking the blind spot.

Structured data	
Weather	?
Contributing factor	Lane change
Light Condition	Daylight

Crash 3

A Chevrolet Malibu collided head-on with a Subaru Outback on a rural road during foggy conditions at night.

Structured data	
Weather	Foggy
Contributing factor	?
Light Condition	Dark

Inconsistent data, missing values for varying attributes

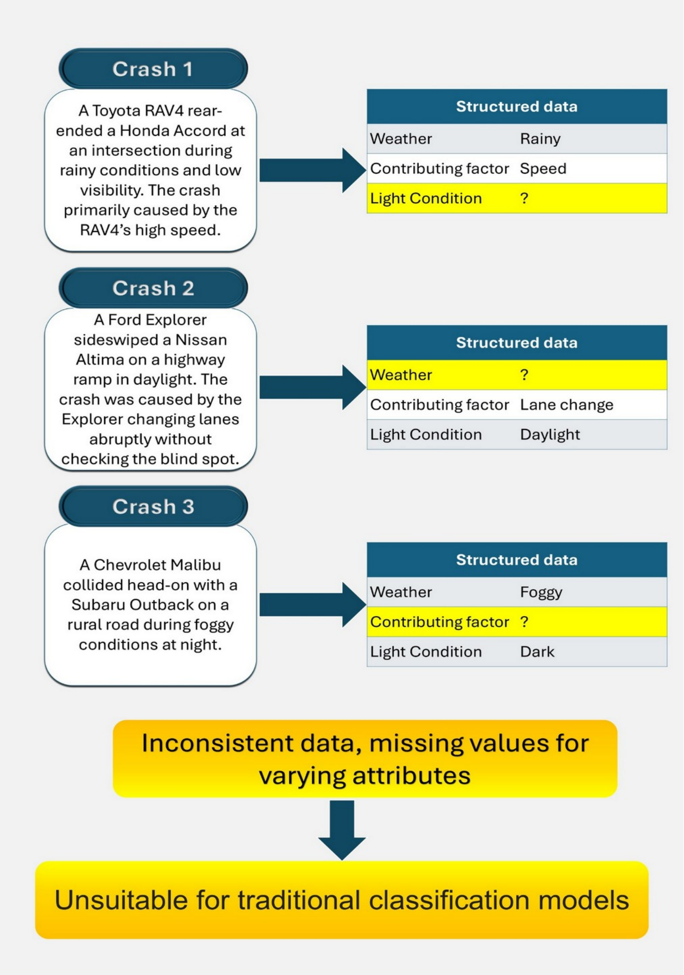
Unsuitable for traditional classification models

New Jersey Institute of Technology

Intelligent Transportation Systems Resource Center

Background

- Challenges in Crash Severity Prediction
  - **Crash Report Structure:** Crash reports typically contain a combination of structured, descriptive crash features (parameters), and crash narratives.
  - **Incomplete or missing data** in both structured and narrative fields
  - **Inconsistent formats**
  - **Narratives vary in wording and detail**



Crash 1

A Toyota RAV4 rear-ended a Honda Accord at an intersection during rainy conditions and low visibility. The crash primarily caused by the RAV4's high speed.

Structured data	
Weather	Rainy
Contributing factor	Speed
Light Condition	?

Crash 2

A Ford Explorer sideswiped a Nissan Altima on a highway ramp in daylight. The crash was caused by the Explorer changing lanes abruptly without checking the blind spot.

Structured data	
Weather	?
Contributing factor	Lane change
Light Condition	Daylight

Crash 3

A Chevrolet Malibu collided head-on with a Subaru Outback on a rural road during foggy conditions at night.

Structured data	
Weather	Foggy
Contributing factor	?
Light Condition	Dark

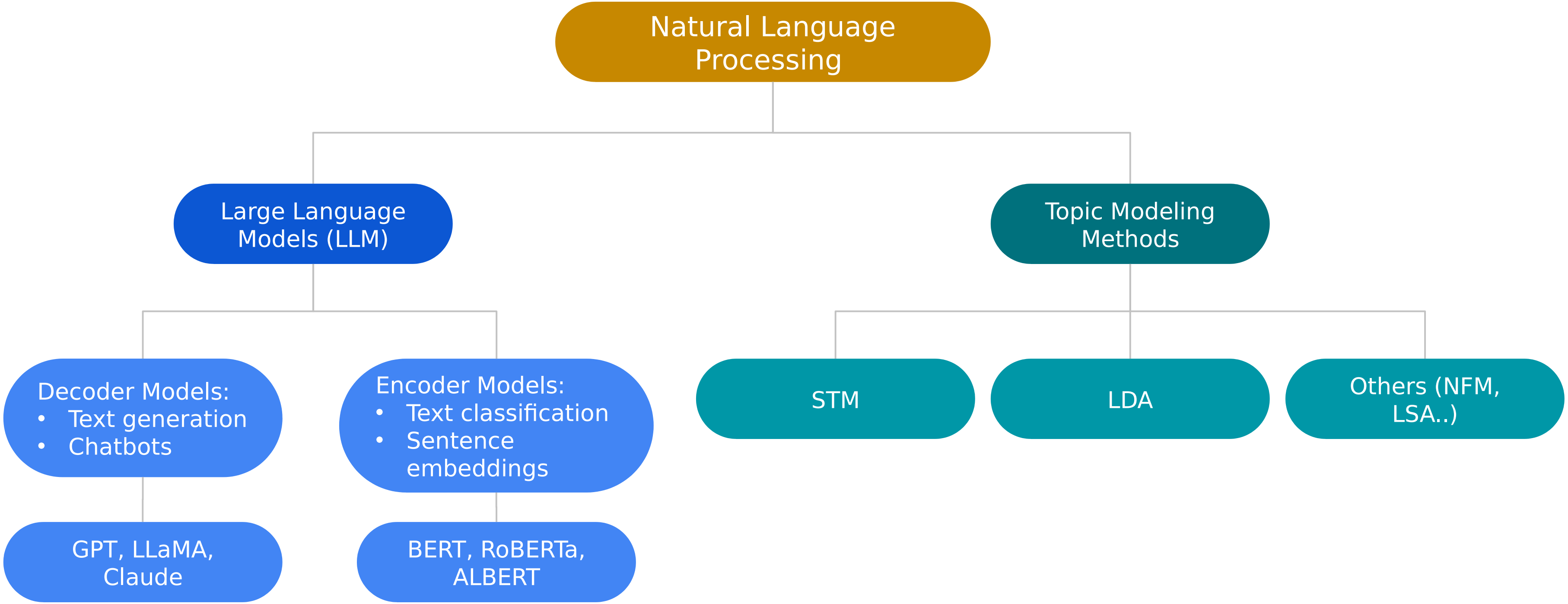
Inconsistent data, missing values for varying attributes

Unsuitable for traditional classification models

Research Objective

- **Develop a crash severity prediction model** that leverages both structured data and synthesized crash narratives
- **Generate consistent, informative crash descriptions** by transforming structured parameters into synthetic narratives
- **Apply large language models (LLMs)** to analyze and predict crash severity from these synthesized narratives

Methodology & Data



Natural Language Processing

Large Language Models (LLM)

Decoder Models:
- Text generation
- Chatbots

GPT, LLaMA, Claude

Encoder Models:
- Text classification
- Sentence embeddings

BERT, RoBERTa, ALBERT

Topic Modeling Methods

STM

LDA

Others (NFM, LSA..)

NJIT New Jersey Institute of Technology

ITSRC Intelligent Transportation Systems Resource Center

The Model Dataset

- 3,293,688 crash records from January 2010 to November 2022, tabulated.
    - From features related to crashes, only the information **about crash time, date, geographic location, and environmental conditions** was utilized.
    - The KABCO crash severity scale was transformed into three categories: 'No Injury,' 'Injury,' and 'Fatal.’
        - No Apparent Injury: **No Apparent Injury**
        - Possible Injury, Suspected Minor Injury: **Injury**
        - Suspected Serious Injury, Fatal Injury: **Serious injury or Fatal**

The Model Dataset

Total	Last 3 Years
• 3293676 Records o No Apparent Injury 2569464 (%78) o Possible Injury 545166 (%16.5) o Suspected Minor Injury 153948 (%4.6) o Suspected Serious Injury 18149 (%0.5) o Fatal Injury 6949 (%0.2)	• 479302 records (%14.5 of total) o No Apparent Injury 375173 (%78.2) o Possible Injury 57442 (%12) o Suspected Minor Injury 39420 (%8.2) o Suspected Serious Injury 5948 (%1.2) o Fatal Injury 1319 (%0.27)

NJIT New Jersey Institute of Technology

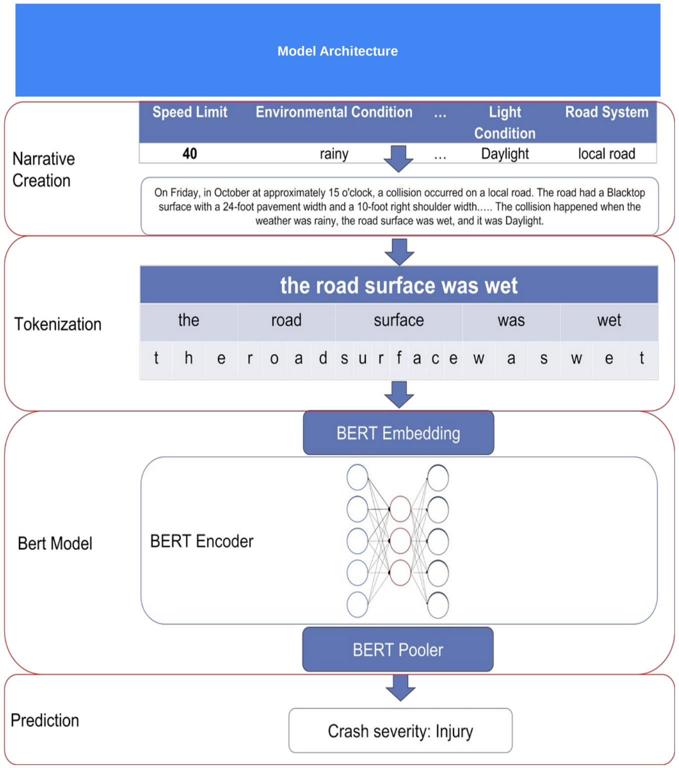
ITSRC Intelligent Transportation Systems Resource Center

Methodology: BERT

Narrative creation: The narratives include six sentences

- The first and second sentence provide the time, date, and information regarding the crash location
- The third and fourth sentence ascertain the traffic information – speed and AADT.
- The fifth sentence ascertains the weather conditions
- The last sentence describes traffic control facilities present at the crash location,

○ **Processing pipeline:** Narratives are tokenized, embedded, and passed through BERT's transformer encoders and pooler to generate contextualized representations for severity prediction.



Speed Limit	Environmental Condition	...	Light Condition	Road System
40	rainy	...	Daylight	local road

On Friday, in October at approximately 15 o'clock, a collision occurred on a local road. The road had a Blacktop surface with a 24-foot pavement width and a 10-foot right shoulder width..... The collision happened when the weather was rainy, the road surface was wet, and it was Daylight.

the road surface was wet

the | road | surface | was | wet

t h e r o a d s u r f a c e w a s w e t

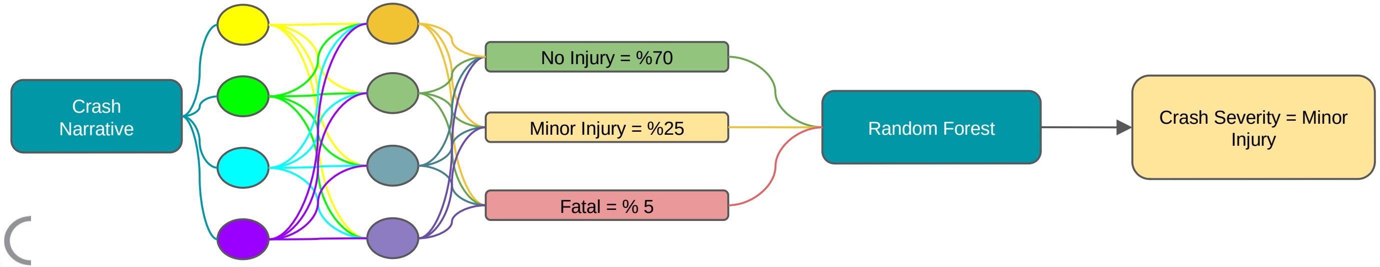
Crash severity: Injury

NJIT New Jersey Institute of Technology

ITSRC Intelligent Transportation Systems Resource Center

Methodology: BERT/Random Forest Hybrid Model

- Class Imbalance
  - Majority of records are **“no injury”**, leading to biased predictions in standard classifiers
  - Even when a crash is severe, its probability may be lower than “no injury” due to imbalance
- Tokenization & Embedding with BERT
  - Use BERT to tokenize crash narratives
  - Generate class probability scores (e.g., for no injury, minor injury, serious injury, fatal)
- Post-BERT Classification with Random Forest
  - Use BERT’s output probability distribution
  - Train a Random Forest classifier on these probability vectors to improve classification of minority classes



Crash Narrative

No Injury = %70

Minor Injury = %25

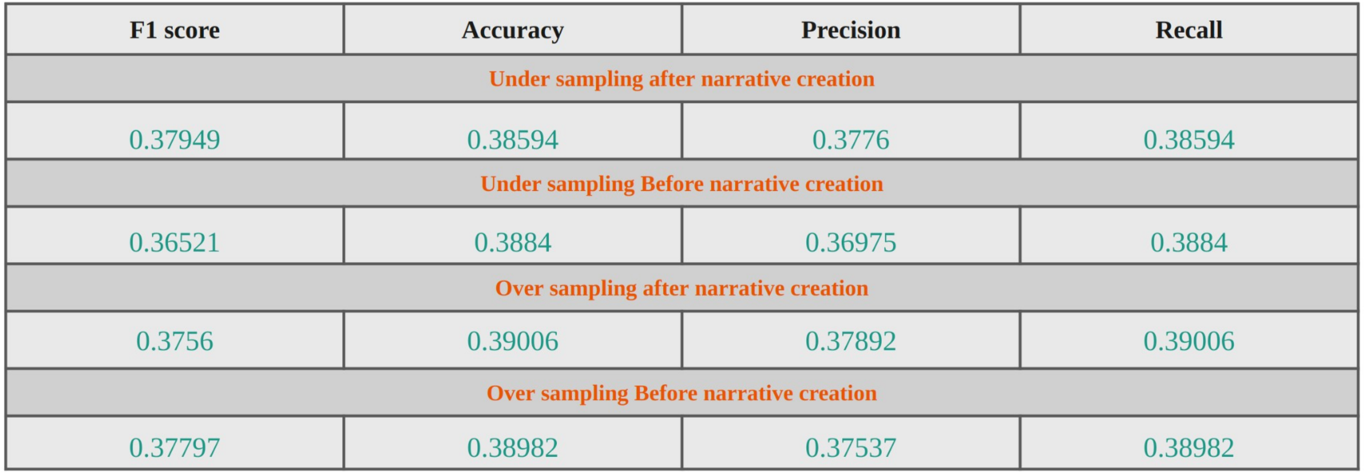
Fatal = % 5

Random Forest

Crash Severity = Minor Injury

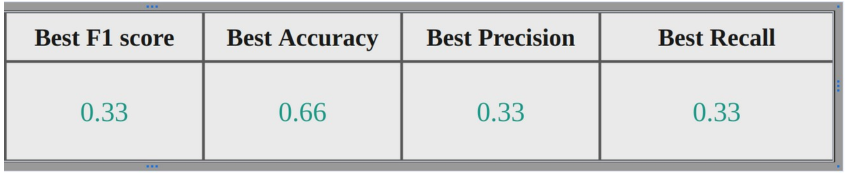
Preliminary Results

- Bert:



F1 score	Accuracy	Precision	Recall
Under sampling after narrative creation			
0.37949	0.38594	0.3776	0.38594
Under sampling Before narrative creation			
0.36521	0.3884	0.36975	0.3884
Over sampling after narrative creation			
0.3756	0.39006	0.37892	0.39006
Over sampling Before narrative creation			
0.37797	0.38982	0.37537	0.38982

- Hybrid model



Best F1 score	Best Accuracy	Best Precision	Best Recall
0.33	0.66	0.33	0.33

What’s Ahead

- Integrate Spatial Imagery
  - Extract location-specific features from satellite or street-level images to generate contextual descriptions (e.g., "intersection near a school zone")
- Incorporate Land Use and Environmental Data
  - Use land use type (e.g., residential, commercial) and surrounding features (e.g., parks, railroads) to enrich narrative content
- Leverage Decoder-Based Language Models
  - Use decoder architectures (e.g., GPT-style models) to generate more realistic and diverse narratives in natural language

Thank You